Course Notes: Deep Learning for Visual Computing

Peter Wonka

October 24, 2021

Contents

1 Information Theory

1.1 Literature

- Literature: Cover and Thomas, Elements of Information Theory

1.2 Videos

- https://youtu.be/bkLHszLlH34

1.3 Relation to other areas

- Information theory has links to
  - physics (thermodynamics, statistical mechanics)
  - computer science (Kolmogorov complexity or algorithmic complexity: complexity of a string of data can be defined by the length of the **shortest** binary **computer program** for computing the string)
  - communication
  - probability and statistics
  - machine learning
  - economics

1.4 Notation

- X: discrete random variable
- $\mathcal{X}$: alphabet
- $p(x) = Pr\{X = x\}, x \in \mathcal{X}$: probability mass function
  - $p(x)$ is short for $p_X(x)$
  - $p(x)$ and $p(y)$ refer to two different random variables (as this stands for $p_X(x)$ and $p_Y(y)$)
- log: base 2 logarithm

1.5 Information Content Design Goals

- Likely events should have low information
- Less likely events should have higher information
- Independent events should have additive information
- Note: intuitively information measures **surprise**

1.6 Information Content

- Definition: **Information Content** of an event

$$I(E) = \log_2 \frac{1}{Pr(E)} = -\log_2 Pr(E) \tag{1.1}$$

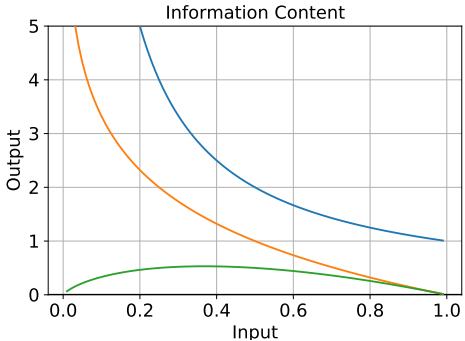
  - E: event
  - $Pr(E)$: probability of the event E

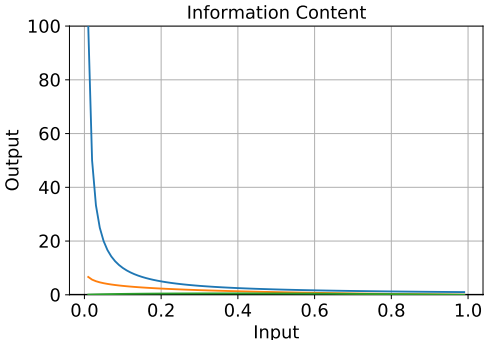
- Definition: Information content for outcome of a random variable $X = x$)

$$I_X(x) = -\log_2 p_X(x) \tag{1.2}$$

  - X: discrete random variable
  - $p_X(x) = Pr\{X = x\}$: probability mass function

1.7 Information Content Graph





- Graph of I(x)
  - blue: $1/x$
  - orange: $log_2(1/x) = -log_2(x)$
  - green: $-x * log_2(x)$

1.8 Information Content Table

x	$1/x$	$-log_2(x)$	$-x\log_2(x)$
0.01	100	6.643856	0.066439
0.1	10	3.321928	0.332193
0.2	5	2.321928	0.464386
0.3	3.333333	1.736966	0.52109
0.4	2.5	1.321928	0.528771
0.5	2	1	0.5
0.6	1.666667	0.736966	0.442179
0.7	1.428571	0.514573	0.360201
0.8	1.25	0.321928	0.257542
0.9	1.111111	0.152003	0.136803
1	1	0	0

1.9 Information Content Example

- Comparing the information content of weather events in Austria and Saudi Arabia

Probability	Sunny	Cloudy	Rainy
Saudi Arabia	0.9	0.09	0.01
Austria	0.4	0.3	0.3

Information	Sunny	Cloudy	Rainy
Saudi Arabia	0.15	3.47	6.64
Austria	1.32	1.74	1.74

1.10 Entropy

- Entropy of discrete random variable X:

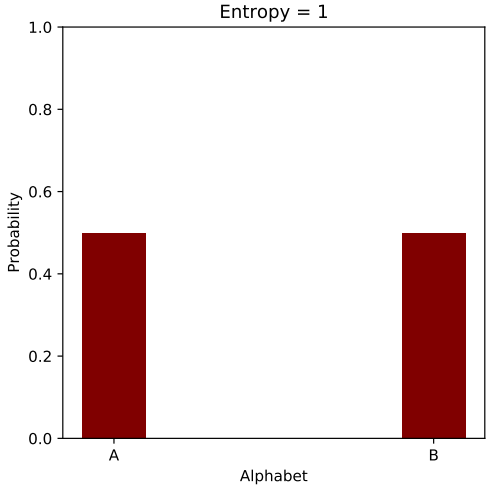
$$H(X) = \mathbb{E}\big(I_X\big) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{1.3}$$

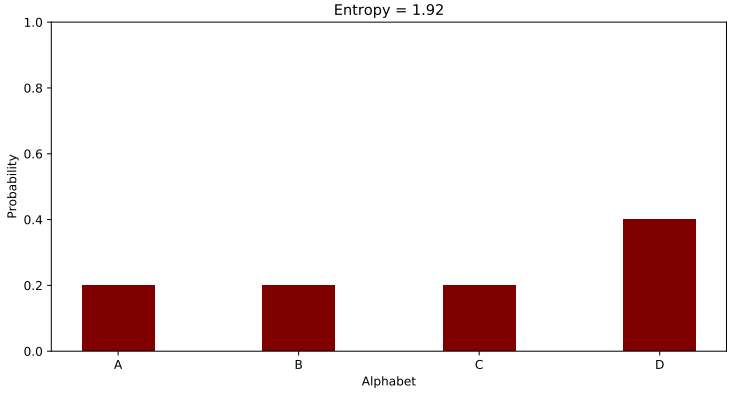
  - measures the **average uncertainty**
  - How many bits on average are required to describe an outcome of the random variable?
  - $0log(0) = 0$ and more generally $0log(x) = 0$

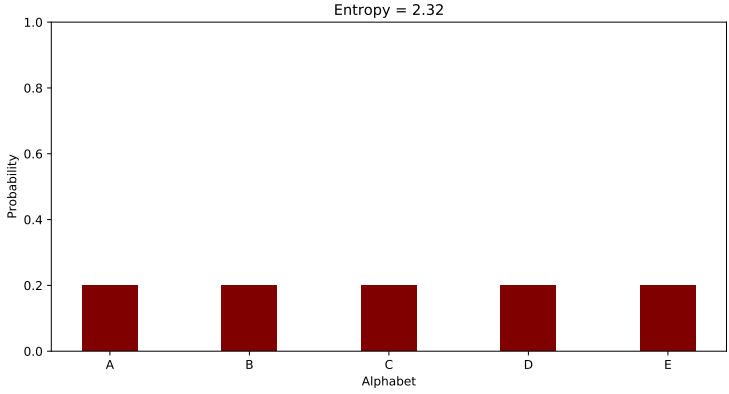
1.11 Entropy Example

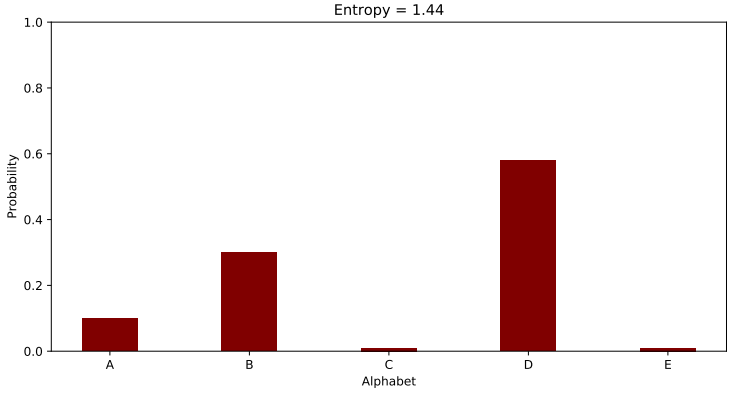
- $H(X) = 1.57 > 0.516 = H(Y)$
  - X is a random variable according to the weather in Austria
  - Y is a random variable according to the weather in Saudi Arabia
- Tomorrows weather in Saudi Arabia is more certain (very likely sunny)
- Tomorrows weather in Austria is more uncertain (probabilities are almost equal)

1.12 Entropy Graphs




Entropy = 1.92
1.0
0.8
0.6
0.4
0.2
0.0
Probability
A
B
C
D
Alphabet


Entropy = 2.32
1.0
0.8
0.6
0.4
0.2
0.0
Probability
A
B
C
D
E
Alphabet


Entropy = 1.44
1.0
0.8
0.6
0.4
0.2
0.0
Probability
A
B
C
D
E
Alphabet

1.13 Entropy Comments

- Instead of $H(X)$ we can also write $H(p)$
- log with basis 2 leads to **bits**, the natural log leads to **nats**
- Entropy is highest for a uniform probability mass function

1.14 Entropy Facts

- $H(X) \geq 0$
- $H(X) \leq log|\mathcal{X}|$

1.15 Joint Entropy

- **Joint entropy** $H(X, Y)$ of a pair of random variables:

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \tag{1.4}$$

- this follows directly from the definition of entropy

1.16 Joint Entropy Discussion

- Joint entropy can also be defined for n random variables $H(X_1, X_2, ..., X_n)$

1.17 Joint Entropy Facts

- $H(X, Y) \geq 0$
- $H(X, Y) \geq \max(H(X), H(Y))$
- $H(X_1, X_2, ..., X_n) \geq \max_i (H(X_i))$
- $H(X_1, X_2, ..., X_n) \leq H(X_1) + H(X_2) + \ldots + H(X_n)$

1.18 Conditional Entropy

- If $(X, Y) \sim p(x, y)$, the **conditional entropy** $H(Y|X)$ is defined:

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \tag{1.5}$$

$$= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \tag{1.6}$$

$$= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)} \tag{1.7}$$

$$= -\mathbb{E}\big(\log p(Y|X)\big) \tag{1.8}$$

1.19 Conditional Entropy Facts

- $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
- $H(Y|X) \neq H(X|Y)$
- $H(X) - H(X|Y) = H(Y) - H(Y|X)$
- $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$

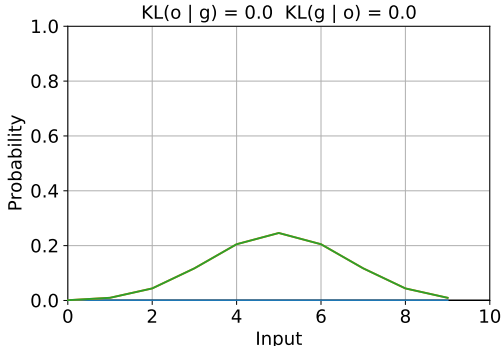
1.20 Relative Entropy = KL-Divergence

- The **relative entropy** or **KL-divergence** can measure the difference between two probability distributions

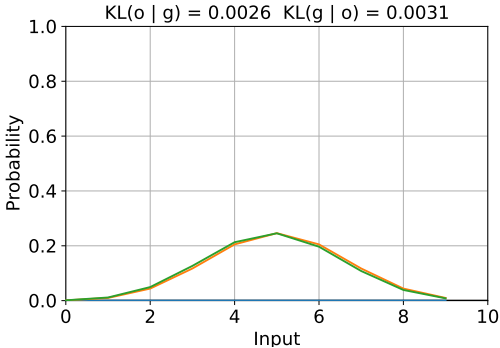
$$D_{KL}(p \,||\, q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \tag{1.9}$$

- $0 \log \frac{0}{q} = 0$
- $p \log \frac{p}{0} = \infty$
- $0 \log \frac{0}{0} = 0$
- If there is any symbol $x \in \mathcal{X}, p(x) > 0, q(x) = 0$ then $D_{KL}(p \,||\, q) = \infty$
- relative entropy is a measure for the inefficiency of assuming that the distribution is q when the true distribution is p
- If we construct a code with distribution q instead of true distribution p we need $H(p) + D_{KL}(p \,||\, q)$ bits on average

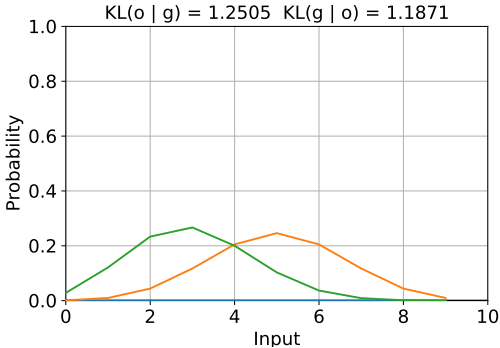
1.21 KL-Divergence Graphs

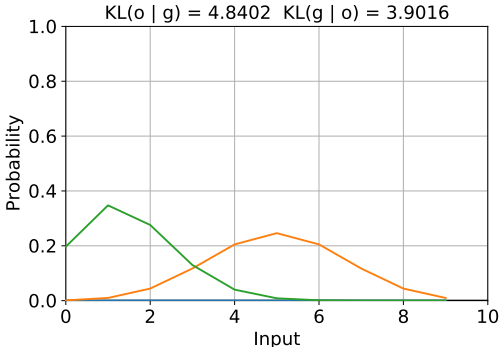


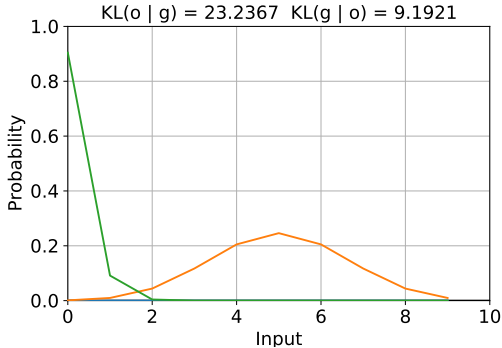
- o stands for orange, g stands for green
- two binomial distributions



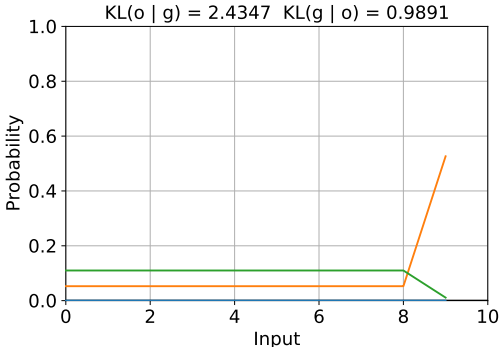
- slight shift


KL(o | g) = 1.2505 KL(g | o) = 1.1871
1.0
0.8
0.6
0.4
0.2
0.0
Probability
0
2
4
6
8
10
Input

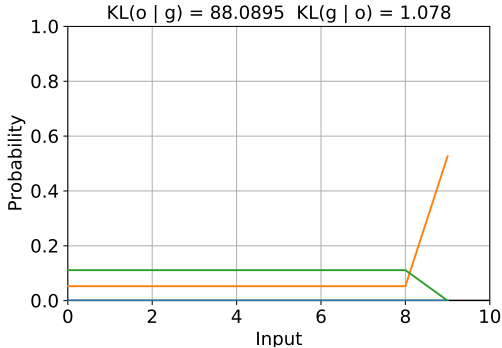

KL(o | g) = 4.8402 KL(g | o) = 3.9016
1.0
0.8
0.6
0.4
0.2
0.0
Probability
0
2
4
6
8
10
Input



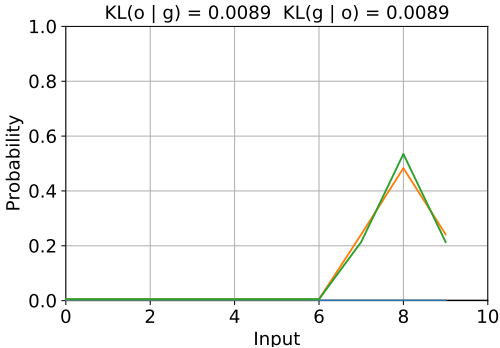
- KL-divergence between orange and green is very high because green probabilities for high values, e.g. 8, 9, 10, are very low.

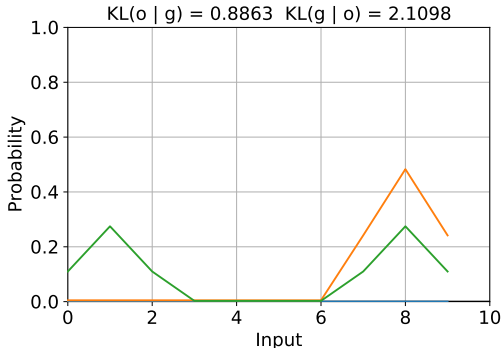


- two uniform distributions, except for the last value



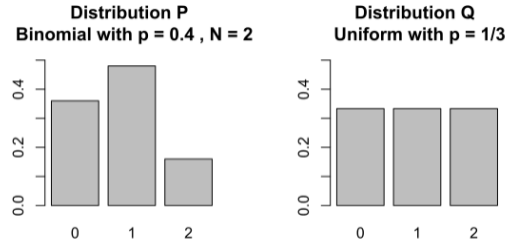
- last value for green got a lot smaller $1e-50$
- an important factor are very small values in the second distribution


KL(o | g) = 0.0089 KL(g | o) = 0.0089
1.0
0.8
0.6
0.4
0.2
0.0
Probability
0
2
4
6
8
10
Input



- If the second distribution has an extra bump, it's not that bad.
- If the second distribution is missing a bump, it's worse.

1.22 KL-Divergence Example



- Two distributions:

x	0	1	2
P(x)	0.36	0.48	0.16
Q(x)	0.333	0.333	0.333

- Computation:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \ln\left(\frac{P(x)}{Q(x)}\right) \tag{1.10}$$

$$= 0.36 \log\left(\frac{0.36}{0.333}\right) + 0.48 \log\left(\frac{0.48}{0.333}\right) + 0.16 \log\left(\frac{0.16}{0.333}\right) \tag{1.11}$$

$$= 0.012032552 + 0.076013996 - 0.051001402 = 0.037045146 \tag{1.12}$$

$$D_{\mathrm{KL}}(Q \parallel P) = \sum_{x \in \mathcal{X}} Q(x) \ln\left(\frac{Q(x)}{P(x)}\right) \tag{1.13}$$

$$= 0.333 \log\left(\frac{0.333}{0.36}\right) + 0.333 \log\left(\frac{0.333}{0.48}\right) + 0.333 \log\left(\frac{0.333}{0.16}\right) \tag{1.14}$$

$$= -0.011141252 - 0.052787497 + 0.106252921 = 0.042324172 \tag{1.15}$$

- Two distributions (changed):

x	0	1	2
P(x)	0.36	0.48	0.16
Q(x)	0.99	0.009	0.001

$$D_{\mathrm{KL}}(P \parallel Q) = -0.15815977 + 0.828959389 + 0.352659197 = 1.023458817 \tag{1.16}$$

$$D_{\mathrm{KL}}(Q \parallel P) = 0.434939367 - 0.015542989 - 0.00220412 \quad = 0.417192258 \tag{1.17}$$

1.23 Mutual Information

- **Mutual Information** is the relative entropy (KL-divergence) between the joint probability $p(x, y)$ and the product of the marginal probability mass functions $p(x)$ and $p(y)$:
- **Mutual Information** is defined as:

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{1.18}$$

1.24 Mutual Information Discussion

- Mutual information $I(X;Y)$ is the **reduction in** the **uncertainty** of X due to the knowledge of Y.
- Mutual information is symmetric
- Mutual information of a random variable with itself is the entropy of the random variable

1.25 Mutual Information Facts

- $I(X;Y) = I(Y;X)$
- $I(X;Y) = D(p(x,y) \parallel p(x)p(y))$
- $I(X;Y) = H(X) - H(X|Y)$ and $I(X;Y) = H(Y) - H(Y|X)$
  - mutual information is the reduction in uncertainty of X due to the knowledge of Y
- $I(X;Y) = H(X) + H(Y) - H(X,Y)$
- $I(X;X) = H(X) - H(X|X) = H(X)$

1.26 Cross Entropy

- **Cross Entropy** is defined as:

$$H_{cross}(p, q) = H(p, q) = -\sum_{x \in \mathcal{X}} p(x) log_2(q(x)) \tag{1.19}$$

1.27 Cross Entropy Discussion

- Someone had the brilliant idea to use the same notation for joint entropy and cross entropy

1.28 Cross Entropy Classification Example

- Ground truth class probability vector = $(0, 0, 0, 1, 0)$
- Predicted class probabilities by the network = $(0.1, 0.2, 0.1, 0.5, 0.1)$
- Cross Entropy = -$log_2(0.5)$
  - logarithm of the predicted probability for the correct class.